UNIVERSITY OF HELSINKI

Language Choice between English and Finnish

Insights from Geotagged Social Media

Katri Loikkanen
Master's Thesis
Master's Programme in English Studies
University of Helsinki
April 2020

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Humanistinen tiedekunta		Koulutusohjelma – Utbildningsprogram – Degree Programme Englannin kielen ja kirjallisuuden maisteriohjelma
Opintosuunta – Studieinriktning – Study Track Englannin kielen ja kirjallisuuden maisteriohjelma		
Tekijä – Författare – Author Loikkanen, Katri Tuulia		
Työn nimi – Arbetets titel – Title Language Choice between English and Finnish – Insights from Geotagged Social Media		
Työn laji – Arbetets art – Level maisterintutkielma	Aika – Datum – Month and year huhtikuu 2020	Sivumäärä– Sidoantal – Number of pages 43

Tiivistelmä – Referat – Abstract

Tutkielma tarkastelee suomalaisten Instagram-käyttäjien kielivalintoja englannin ja suomen välillä Helsingin Senaatintorilta kerätyssä aineistossa. Tutkielma selvittää, miten erilaiset tekijät kuten julkaisuajankohta ja viestin sisältö vaikuttavat käyttäjien kielivalintoihin.

Tutkielman aineisto koostuu lähes 120 000 Instagram-julkaisusta. Julkaisut on ladattu palveluun ja paikkaleimattu Senaatintorin lähistölle vuosina 2013-2018. Käyttäjille määriteltiin todennäköisin kotimaa sen perusteella, missä he olivat olleet pisimpään aktiivisia Instagram-julkaisujen perusteella. Päivitysten kieli määriteltiin automaattisen kielentunnistuksen avulla. Koska tavoitteena oli tutkia suomalaisten käyttäjien kielivalintoja englannin ja suomen välillä, lopulliseen aineistoon sisältyivät ainoastaan suomalaisten käyttäjien tekemät päivitykset, joiden kieli oli joko englanti tai suomi. Lopullisen aineiston koko oli 16 987 päivitystä.

Tutkimuksen pääasiallinen menetelmä on logistinen regressio, joka on määrällinen tilastollinen metodi, jolla voidaan mallintaa kaksiarvoisen muuttujan todennäköisyyksiä usean riippumattoman muuttujan avulla. Regressioanalyysin tulokset ilmaistaan ns. riskeinä (engl. ‘odds’). Tässä tutkimuksessa riippuva, kaksiarvoinen muuttuja oli valinta englannin ja suomen välillä. Riippumattomat muuttujat, joiden avulla riskit määriteltiin, liittyivät päivitysten julkaisuajankohtaan ja siihen, sisältyikö päivitykseen maininta sijainnista tai tietyistä tapahtumista.

Logistinen regressio toteutettiin Python-ohjelmointikielen avulla. Tieto jokaisen päivityksen julkaisuajankohdasta oli valmiiksi kirjattuna aineistoon, ja aikaan liittyvät muuttujat luotiin näiden tietojen perusteella. Päivitysten sisältöön liittyvät muuttujat taas määriteltiin sisältöanalyysin avulla. Analysoin ensin AntConc-sovelluksen avulla aineistossa useimmin esiintyviä sanoja. Seuraavaksi otin aineistosta kaksi satunnaisotosta ja analysoin näiden otosten sisältöä laadullisesti. Analyysi paljasti kaksi aineistossa usein toistuvaa aihetta: julkaisun sijainti (Senaatintori, Helsinki tai Suomi) sekä Senaatintorilla järjestettävät tapahtumat ja juhlat (esim. Lux Helsinki ja joulu). Näistä teemoista tehtiin muuttujia tarkistamalla, esiintyykö julkaisuissa tiettyjä näihin aiheisiin liittyviä avainsanoja.

Tutkimuksen tulokset osoittavat, että Suomen mainitseminen Instagram-julkaisussa kasvattaa riskiä valita kieleksi englanti huomattavasti. Yksi selitys tälle yllättävälle tulokselle saattaa olla se, että kansainväliselle yleisölle suunnatussa julkaisussa tarvitaan todennäköisemmin erillinen maininta Suomesta. Myös Helsingin mainitseminen kasvattaa riskiä valita englanti jonkin verran. Sen sijaan tapahtumien tai juhlien mainitseminen kasvattaa riskiä valita suomi. Nämä tapahtumat saatetaan kokea erityisen paikallisina tai suomalaisina. Julkaisuajankohdalla sen sijaan ei näyttäisi olevan mitään vaikutusta kielivalintaan. Hypoteesini oli, että englantia käytetään enemmän iltaisin, viikonloppuisin ja kesäisin, mutta tulokset eivät tue tätä. Vaihtoehtoinen regressiomalli, joka ei sisältänyt ajallisia muuttujia, ei muuttanut sisältöön liittyvien muuttujien tuloksia. Tulos tukee sitä, että julkaisuajankohta ei vaikuta kielivalintaan.

Tutkielma osoittaa, että logistisen regression kaltaiset tilastolliset menetelmät soveltuvat erinomaisesti kielivalintojen tutkimiseen. Regressioanalyysin ja laadullisten menetelmien yhdistelmä tuo uusia näkökulmia kielivalintoihin vaikuttaviin tekijöihin. Tutkimassani aineistossa viestin sisältö vaikuttaa kielivalintoihin huomattavasti enemmän kuin julkaisuajankohta. Tuloksia tulkitessa tulee kuitenkin ottaa huomioon, että aineisto sijoittuu maantieteellisesti tarkasti rajatulle alueelle, joten tuloksia ei välttämättä voi yleistää. Tarvitaan siis lisää vastaavanlaisia tutkimuksia muista aineistoista selvittämään, pätevätkö tutkimuksen tulokset yleisesti.

Avainsanat – Nyckelord – Keywords
kielivalinta, monikielisyys, englanti, suomi, sosiaalinen media, Instagram, logistinen regressio

Säilytyspaikka – Förvaringställe – Where deposited
Helsingin yliopiston kirjasto

Muita tietoja – Övriga uppgifter – Additional information

Contents

1 Introduction

The English language plays an increasingly important role in Finnish society and culture. A large-scale survey into the uses of and attitudes towards English in Finland (Leppänen et al., 2011) revealed that Finns, especially younger generations, consider English an integral part of their linguistic repertoire. Previous studies (e.g. Taavitsainen and Pahta, 2003; Leppänen, 2007; Leppänen and Nikula, 2007; Taavitsainen and Pahta, 2008; Leppänen et al., 2009; Piirainen-Marsh, 2010; Kääntä et al., 2013; Laitinen, 2014) have looked into the functions of English in widely varying social contexts such as business, education, youth culture and video games. These studies have noted that Finnish people use more and more English in their daily lives and that English has acquired a wide variety of social and symbolic meanings.

A more recent development not addressed by these studies is the emergence and increased popularity of social media. Social media platforms offer especially interesting opportunities for the study of English in Finland since a large portion of communication on these platforms happens in English, and users of social media tend to be mostly young people, i.e. the people who use English most frequently (Leppänen et al., 2011). At the moment, however, there is still very little research on this topic, and one of the main motivations for this MA thesis is to begin filling this gap.

A great opportunity for this is offered by interesting new data collected by Hiippala et al. (2019). The data consist of nearly 120 000 Instagram posts made in the close vicinity of Helsinki Senate Square over a period of approximately four and a half years. Hiippala et al. found that English is by far the most dominant language in the "virtual linguistic landscape" (2019, pp. 291) of the Senate Square. Interestingly, they also found that, at least in this specific location, Finnish Instagram users write approximately half of their posts in English. The aim of this MA thesis is to investigate some of the factors that might affect language choice between Finnish and English among Finnish Instagram users. I use qualitative content analysis to determine which factors to include in my analysis, and then use logistic regression to

investigate the effect of these factors on language choice. In other words, this MA thesis is a mixed methods study.

In Chapter 2, I first address some terminological considerations, followed by an account of previous research on English in Finland as well as previous research on language choice on social media, and finally a brief look into the connections between multilingualism, online language use and place. In Chapter 3, I describe my data set which is a large collection of Instagram posts collected from Helsinki Senate Square. In Chapter 4, I first outline the primary method of this study, logistic regression. I then explain how I chose the variables used in the logistic regression model; some of them were based on metadata provided by Instagram, and some were chosen based on qualitative content analysis of the data set. In Chapter 5, I present and analyze the results of the logistic regression. In Chapter 6, I summarize my main results, compare them to earlier research, point out some of the limitations of this study as well as suggest topics for further research. Chapter 7 concludes the thesis.

2 Background

2.1 Terminological considerations: SNS vs. social media

Before diving into previous research, I would briefly like to address some terminological concerns that come up when discussing language use online. Many different, partially overlapping terms exist to refer to platforms like Instagram: 'social media', 'social networking', 'Web 2.0', 'the social web', and so on. What makes their usage tricky is that they have no strictly defined, established meanings and are used slightly differently in different contexts and by different research communities. For example, it is sometimes difficult to say which websites/platforms count as examples of social media and which do not. In addition, these terms seem to have slightly different meanings and usages in everyday colloquial speech and in linguistic research.

McCay-Peet and Quan-Haase (2017) define the term social media as "web-based services that allow individuals, communities, and organizations to collaborate, connect, interact, and build community by enabling them to create, co-create, modifies [sic], share, and engage with user-generated content that is easily

accessible" (pp. 17). The consensus among researchers seems to be that social media is a broader umbrella term for "Internet-based sites and services that promote social interaction between participants" (Page et al., 2014, pp. 5; see also Zappavigna, 2012). According to this definition, social media refers not only to the likes of Facebook and Twitter, but also to blogs, wikis, discussion forums and so on – any platform that enables social interaction or collaboration. Social network(ing) or social network(ing) site/service (hereafter SNS), on the other hand, refers to a specific type of social media platform that "reframe[s] the dialogic links between participants as a network, increasing the number, visibility and reach of an individual's connections with others in online spaces" (Page et al., 2014, pp. 7). Examples of SNSs include popular platforms such as Facebook and Twitter, as well as platforms which used to be popular but are now largely abandoned or closed down completely, such as Myspace, Google+ and Bebo. There is some ambiguity as to whether Instagram is an example of SNSs. Although most researchers seem to categorize it as such, according to McCay-Peet and Quan-Haase's (2017) classification, it falls into the category of media sharing. I contend that even though the sharing of photographs is usually considered the main function of Instagram, it is also a clear example of an SNS because of the social networks formed through "following" and being followed by specific users.

As well as being a more specific term than social media, SNS also seems to be the preferred choice in much of previous research on language choice on e.g. Facebook (see Section 2.3 for examples of these studies). While social media seems to be the preferred term in non-academic discourse and therefore might be more familiar to a wider audience, for the reasons outlined above, I am going to use SNS when referring to Instagram and other similar platforms, and social media as a broader term.

2.2 Previous Research on English in Finland

The role of English in Finland is undergoing major changes. Traditionally, English has been seen as foreign language that is used primarily for international communication. However, it seems that English is increasingly being used also in intranational and intracultural contexts, alongside or sometimes instead of the national languages, Finnish and Swedish (see e.g. Leppänen and Nikula, 2007;

Taavitsainen and Pahta, 2008). The increasing use of English has been a frequent topic of often heated public discussion, with some even seeing it as a threat to the survival and purity of the Finnish language.[1]

A number of studies during the last two decades have taken an interest in the changing roles and usages of English in Finland. Taavitsainen and Pahta (2003) note the increased presence of English in the daily lives of Finnish people due to globalization and through for example different media such as TV (foreign programs are subtitled instead of dubbed), as well as business contexts such as the names of companies and job advertisements which often include English. They argue that, at least for some people, English might be moving from a foreign language to a second language. They also see English as somewhat of a threat to the Finnish language, especially in education, research and business, and argue that there is a danger of domain loss in these areas. They note a trend in other parts of the world of English moving from a foreign language to a second language and then to a first language and seem to imply that there might also be a risk of this happening in Finland. In a later article, Taavitsainen and Pahta (2008) take a rather more nuanced view of the issue, moving away from viewing English as a threat and instead examining the complex roles of English in e.g. institutional contexts, public spaces and media discourse. They also note that the traditional model of English as first language (L1), second language (L2) and foreign language (EFL) which they used in their earlier article no longer makes sense in a linguistic and cultural context such as Finland. They argue that the relationships between English and globalization are more complex than had previously been thought and note that English has taken on local uses and meanings in Finland. They also suggest that English is becoming "a natural part of language resources for Finns" (Taavitsainen and Pahta, 2008, pp. 37).

Leppänen and Nikula (2007) take an in-depth look into English in different media, educational and business contexts in Finland. They argue that the growing influence of English should not be considered a threat to Finnish, but rather a sign of Finns' bilingual and multilingual competence. They see English as a valuable social asset and an important additional linguistic resource that can be used to index cultural

[1] A recent example of this is an article (Heikkinen, 2018) published on the website of *Helsingin Sanomat*, the biggest newspaper in Finland. The article states outright that English is a threat to Finnish and that something must be done about this. Heikkinen cites several concerned politicians and not a single linguist or other expert.

meanings. Like Taavitsainen and Pahta (2008), Leppänen and Nikula also argue that English is becoming more domesticated and a natural part of people's language use and suggest that the strict lines between Finnish and English might be blurring. However, they still view the contrast between the two languages as important (at least in some contexts) and point out that different languages can serve different functions. An important point made by Leppänen and Nikula is that the increasing use of English should not be seen as "one-directional process of English taking over Finnish society" but rather "a process in which English is taken up and made use of by Finns in a variety of ways, in order to serve their own purposes" and that "instead of arguing that Finns are/will be forced to use English […] [they] are/will become increasingly aware of the roles and functions of more than one language in their lives, and […] be able to select, switch between and make use of the languages and their variant styles in ways that are appropriate in the situations, settings and discourses at hand" (2007, pp. 368-369).

In part motivated by these studies, a large-scale national survey into the uses and meanings of, as well as attitudes towards, English in Finland was carried out in 2007 (Leppänen et al., 2011). This survey confirmed that English is indeed an important part of Finnish people's daily lives. English was overwhelmingly the most common foreign language studied, used and encountered by Finns. Almost 60% of respondents considered English moderately important or very important to them personally. Attitudes towards English were mostly positive: English skills were considered important and English was not seen as a threat to the Finnish language. Unsurprisingly, English was considered most important and used most often by respondents who were young, highly educated and living in urban areas.

The survey also confirmed that English is becoming a more natural part of Finnish people's language use. 26.4% of respondents chose the option "using English is as natural to me as using my mother tongue". When asked about the reasons for using English, more than a third of respondents indicated that they use it "for the fun of it" at least once a week. Code switching was frequent in spoken language and attitudes towards it were mostly positive. When asked about reasons for mixing English and the mother tongue (Finnish or Swedish depending on the respondent) in spoken interaction, 76.4% indicated they do not even notice they are doing it, which would

suggest that using English comes very naturally to them. The corresponding figure for written language was 47.5%.

The majority of respondents used most English in their free time. Interestingly, most of the English read and written during respondents' free time was in online contexts such as web pages and e-mails. Searching for information was overwhelmingly the most popular English-language activity online. Other activities included reading newspapers, following discussion forums and playing games.

Seeing as English in Finland is often associated with youth culture, media and especially the Internet, it is no surprise that many qualitative studies have focused on these contexts. These include studies on online language use (Leppänen, 2007; Leppänen et al., 2009), video games (Leppänen, 2007; Leppänen et al., 2009, Piirainen-Marsh 2010) and reality TV (Kääntä et al., 2013). However, there have been very important developments in digital media after these studies were published. Mobile technologies such as smartphones and tablets have become increasingly sophisticated and extremely popular, allowing most people in Finland to access the Internet whenever and wherever they want. SNSs such as Facebook, Twitter and Instagram have also exploded in popularity during the last decade. These changes have no doubt had a considerable effect on Finnish people's use of English online, but at the time of writing, there have unfortunately been no studies on this topic. The present study is an attempt to start bridging this gap.

2.3 Previous Research on Language Choice in SNSs

Unlike Finnish people's language practices on SNSs, language choice and the role of English in online communication in general has already been the topic of several studies. My focus in this section is on studies that deal with SNSs in particular, although a lot of work has been done on social media platforms in general. For example, Androutsopoulos (2007) looks at language choice and code-switching in online discussion forums used by diasporic communities based in Germany. Leppänen et al. (2009) examine young Finns' language choices in various 'new media' contexts, including online discussion forums. Soler-Adillon and Freixa (2017) investigate the language choices of trilingual students when accessing and contributing to Wikipedia.

In Section 2.3.1, I summarize earlier research on language choice on Facebook, which is by far the most thoroughly researched SNS when it comes to this topic. These studies focus mainly on how the (imagined) audience affects language choice for multilingual users. In section 2.3.2, I discuss earlier research on language choice on Twitter and Instagram, which is unfortunately limited to only a few studies. However, these studies are interesting because they have more similar methods and data to the present study.

2.3.1 Facebook

The overwhelming majority of studies on language choice in SNSs focus on Facebook, one of the most popular SNSs in the world. Practically all of these studies deal to some extent with the question of how Facebook users' awareness of their (imagined) audience affects their language choices. Many of them use the framework of Bell's (1984) *audience design*, which "posits that a speaker's stylistic choices can in great part be shaped by their attempts to accommodate to their addressees and to others present in the exchange" (Tagg and Seargeant, 2014, pp. 161).

Seargeant, Tagg and Ngampramuan (2012) study language choice between Thai and English in a community of native Thai speakers on Facebook. They argue that language choices are shaped on the one hand by the particular platform and its affordances, and on the other by addressivity strategies and audience design. In many cases, there are several separate conversations within one exchange (a Facebook post and its comments), all exhibiting different linguistic choices. This suggests that diverse language choices are largely accepted by this community. However, in some exchanges, language choice seems to index a particular group identity for the community of English-speaking Thai Facebook users, although the relationship between language choice and identity is not a straightforward one. The authors also focus attention on the role of English as a global language. One of the most interesting findings of this study is that people who share a first language (Thai) and who state that they rarely use any English elements in their offline conversations nevertheless use English with each other on Facebook.

Cunliffe, Morris and Prys (2013) look at young Welsh speakers' language choices on Facebook. They argue that language choices may be affected by "the sender, the

intended audience and the message itself" (Cunliffe, Morris and Prys, 2013, pp. 351). Their findings indicate that online language use is to some extent a continuation of offline language use: the languages present in a user's offline community affect the languages used on Facebook. English seems to be dominant and is at least to some extent seen by the participants as the language of the Internet. English is used even by those users who are not confident in their English skills, and those who use both Welsh and English orally tend to use mostly English on Facebook. It should be noted that these results are based only on the participants' self-reporting of their language use and not empirical observations and might therefore not be reliable.

Lee (2014) investigates the language choices of university students in Hong Kong. The languages used by the participants include Cantonese, Chinese, Hakka and English. According to Lee, the most important factors affecting language choice are users' self-evaluations of their language skills, their awareness of different audiences and different aspects of their identity/self-presentation. For example, different languages may be associated with different roles for the participants. The topic and tone of the post may also have an effect on language choice. For example, some participants stated in interviews that they use "standard Chinese for more 'serious' posts such as status updates that express their emotions and opinions towards certain events in their life [and] Cantonese or a mixture of Cantonese and English […] for more 'mundane' activity updates" (Lee, 2014, pp. 100).

Tagg and Seargeant (2014) examine audience design and language choice in the social networks of three multilingual Facebook users. They argue that audience design strategies play a central role for SNS users wanting to "target individuals and communities from within the wider audience" (Tagg and Seargeant, 2014, pp. 161), and that for multilingual users, language choice in particular becomes an important strategy for audience design. Language choices can be used to include or exclude certain people or groups. There is often a distinction between a local, nationally defined community (such as family and local friends who share a first language), and other, non-local groups (such as a wider circle of international friends who do not share a first language and use English as a lingua franca). The authors pay special attention to the role of English, which can sometimes "function as a marker of inclusivity" and "[facilitate] the formation of translocal communities", whereas local

languages may “function as a strategy for addressing a particular language community and/or for making a post more private”, although in practice this division is not always so clear-cut (Tagg and Seargeant, 2014, pp. 181-182). Tagg and Seargeant also acknowledge that there may be other factors affecting language choice, such as topic, setting, technological and social variables as well as speaker biography.

Androutsopoulos (2014) looks at how *context collapse* (a situation in which people from varied backgrounds who would not usually interact with each other are brought together into the same online community) and audience design affect language choices in the social networks of four young German multilinguals. Androutsopoulos argues that in heterogenous, multilingual social networks, language choice “becomes a key resource by which to bring together or separate various parts of the networked audience” (2014, pp. 71). In general, there are inclusive or *maximizing* contributions and exclusive or *partitioning* contributions. Use of a common-denominator language such as English is the most common maximizing strategy, whereas languages such as Greek and Chinese are used to partition the audience. As in many of the studies cited previously, English “emerges in the findings as an important resource for audience design in a community where it is not an everyday spoken language” (Androutsopoulos, 2014, pp. 72). Similar to the findings of Seargeant, Tagg and Ngampramuan (2012), English is not just used in international contexts but also as a local resource between young Germans.

Birnie-Smith (2015) compares language choices on Kaskus, an Indonesian online discussion forum, and Facebook, using SIDE theory (Social Identity model of De-individuation Effects) and audience design. According to SIDE theory, a high degree of anonymity in a group leads to a decreased perception of individual differences and therefore higher group salience and tighter adherence to group norms. On the other hand, identifiability of group members leads to an awareness of individual differences and therefore a focus on personal identity. Kaskus is an anonymous discussion forum with both implicit and explicit rules about language use. As predicted, Kaskus users adhered to these norms: in a thread encouraging the use of Teochew, all participants chose to post in Teochew. Facebook, on the other hand, is not anonymous and has no explicit rules about language use. On Facebook,

participants exhibited more diverse language choices that seemed to highlight their personal identities. Interestingly, several participants used a lot of English even though most of their Facebook friends do not speak it. English therefore had no practical use for the participants and was perhaps used to highlight their identity as "cool", "modern", or "international".

Hinrichs (2016) studies the language choices of German adults on Facebook. According to Hinrichs, Facebook users choose their baseline language based on to whether all their friends understand German or not. All departures from the baseline language in the data are explained by either audience specification or formulaic phrases. For a user whose baseline language is English, German can function to partition the audience. As Hinrichs states: "Overall, the strategy of audience maximization accounts for most of the observed unexpected choices of English in initial contributions, and audience partitioning accounts for most unexpected uses of languages that are neither a participant's baseline choice nor English" (2016, pp. 29). Language choice in initiating posts differs from language choice in responding posts: English is common in initiating but rare in responding. It seems that responding posts are designed for just the person being responded to and not for a larger audience. Local topics are often discussed in a local language, but this is not always the case. Code-switching and other playfulness is very rare in the data, which might be explained by the fact that adults have more stable identities. Surprisingly, English is used mainly for practical reasons and does not seem to have symbolic value.

To summarize, all of these studies highlight the effect of being aware of one's audience on choice of language. Language choices can be used to target certain parts of the overall audience and to exclude others. There are also many other factors affecting language choice, including identity, the topic and tone of the post and features of the SNS platform. Online language use is affected by offline language use to a large extent (Cunliffe, Morris and Prys, 2013), but users also make language choices on Facebook that they would not normally make in offline conversations (Seargeant, Tagg and Ngampramuan, 2012). The role of English emerges as important in all these studies. English is often used to maximize the audience (e.g. Androutsopoulos, 2014) due to its role as an international lingua franca. Sometimes

English is used merely for practical reasons (Hinrichs, 2016), and sometimes it can have a lot of symbolic value (Birnie-Smith, 2015).

While these studies provide very interesting data on the particular language situations and communities being studied, it should be noted that the generalizability of the results of these studies is rather questionable. They all use qualitative analysis only and have relatively small sample sizes. Many of them also rely partly on participants' reflections on their own and others' language use, which might not always be accurate. Cunliffe, Morris and Prys (2013) rely entirely on self-reporting with no empirical analysis of language use (although this is understandable since studying children raises ethical issues). Although most of these studies do mention other factors impacting language choice, the analysis still focuses heavily on the role of the audience. It would be interesting to see studies on language choice on Facebook that use larger sample sizes, quantitative methods and/or examine the effect of other factors such as topic or tone.

2.3.2 Twitter and Instagram

Within the field of linguistics, Twitter and Instagram are both very much under-researched SNSs, especially compared to Facebook. To my knowledge, at the time of writing, there are only a few studies on language choice on Twitter and only one on Instagram. These studies are also very different from the ones on Facebook – they do not use the framework of audience design. I consider these studies more relevant to the present study since they make use of larger data sets and quantitative methods. Some of them also use similar methods as the present study, and one uses data from the same SNS (Instagram).

Eleta and Golbeck's research (2014) is a notable early example of a quantitative approach to the study of language choice on Twitter, a popular SNS/microblogging site. Similar to the present study, the authors use a logistic regression model to study the effect of different factors on the choice between English and other languages. However, unlike in the present study, they focus on how the characteristics of a bilingual user's social network affect the user's language choices. Their data consist of 92 social networks each centered around a single bilingual user (the central node of the network or *ego*). The languages used by the ego are determined by their last 50

posts, and the languages used by the ego's contacts (followers and followings) are determined by their last 30 posts.

Eleta and Golbeck (2014) use two dependent variables in their logistic regression analysis: the degree of English use and the degree of non-English use in the ego's last 50 posts. The independent variables (the factors whose effect on the ego's language choice is modeled using the logistic regression analysis) are three characteristics of the ego's social network: 1) the proportion of English speakers in the network, 2) the proportion of speakers of the most frequent non-English language in the network (the L2 of the network) and 3) the degree of multilingualism in the network.

Eleta and Golbeck's (2014) findings indicate that, unsurprisingly, the proportion of English speakers in the ego's social network is the strongest predictor for the ego's use of English. Likewise, the proportion of L2 speakers in the network is the strongest predictor for the ego's use of the network's L2. The proportion of English speakers in the network has a strong negative correlation to the ego's use of L2. However, the proportion of L2 speakers does not appear to have such a clear negative effect on the ego's English use. This is in line with many of the studies on Facebook (outlined in the previous section), in which the dominance of English is quite clear. The degree of multilingualism does not seem to be a good predictor of language choice. This is perhaps rather surprising, as it would make sense that the presence of several different languages would encourage the use of English as a lingua franca. The authors sum up their results in the following way:

> "The results of the factor analysis suggest that multilingual Twitter users perceive the language composition of their network and interact accordingly. Or on the contrary, the language choices of multilingual users might attract followers of a specific language profile. Most probably, the relation goes both ways, in a self-feeding cycle." (Eleta and Golbeck, 2014, pp. 431)

There have also been two rather more recent quantitative studies on language choice on Twitter. Laitinen et al. (2018) created the Nordic Tweet Stream, a "real-time monitor text corpus of tweets from the Nordic countries" (p. 349). In 2017, the corpus included more that 12 million tweets from Denmark, Finland, Iceland,

Norway and Sweden. Laitinen et al. found that the main language in the corpus was English, which made up 32.9% of the tweets. It was followed by Swedish (25.9%), Finnish (11.5%), Norwegian (5.5%), Danish (4.9%), and Icelandic (1.9%). Other European languages as well as immigrant languages were also found. When the data were divided according to the individual countries, English was in the top two languages in every country.

In a very recent study, Coats (2019) looks at the connections between language choice and gender in a large corpus of Twitter data from the Nordic countries, consisting of almost 24 million tweets. Tweets written in the main official languages of the Nordic countries as well as English made up the vast majority of the data set. Coats found clear gender variation in the language choices of presumed L1 speakers of Nordic languages. In all the Nordic countries, users classified as female based on their name used more English, whereas users classified as male used more local languages. Males were found to be slightly more likely to write tweets in two or more languages, whereas females were more likely to use only one. These findings are in line with previous sociolinguistic studies, according to which females are more likely to use more socially prestigious language varieties such as English.

Coats argues that female users are leading a shift towards English in Nordic people's use of Twitter. While these results are very interesting and no doubt valid on a general level, I would argue that the conception of gender in this study is somewhat problematic. First, people might not use their real name on Twitter and might sometimes even construct a completely fake online presence that does not align with their actual identity. Second, even a person's real name might not be a good indication of their gender – for example, some transgender people, especially before undergoing gender transition, might still have a name associated with the gender they were assigned at birth and not their real, self-identified gender. Third, the conception of gender as a binary of either 'male' or 'female' is problematic as it excludes groups such as intersex and non-binary people. However, I acknowledge that taking these issues into account is difficult in a data set with very limited information related to gender.

As mentioned, language choice on Instagram is still a very under-researched area, even compared to Twitter. In the only previous Instagram study I was able to find,

Lee and Chau (2018) examine "the relationship between expressions of emotion and language choice in hashtags on Instagram revolving around the 2014 Umbrella Movement in Hong Kong". The Umbrella Movement or Umbrella Revolution was a pro-democracy political movement, demanding among other things universal suffrage in Hong Kong. Lee and Chau (2018) collected 700 posts with the Chinese hashtag #雨傘運動 (''Umbrella Movement"). In all, these posts contained 9049 hashtags and 1289 distinct hashtags.

These tags were then coded for pragmatic function and emotion, as well as language (e.g. Cantonese, standard Chinese, English, mixed code). Most tags were factual in nature: out of the 1289 distinct tags, 370 affective tags were identified. 49.4% of the tags were in English, 37.2% in standard Chinese and 7.8% in Cantonese. Even though much lower than English and Chinese, the amount of Cantonese was surprisingly high as it is usually not used in written form. Analysis of the affective tags indicated that Cantonese and standard Chinese are often used for "expressing emotional anticipation and marking political pursuits, as well as anger, hatred, and frustration" (Lee and Chau, 2018). For supporters of the Umbrella Movement, Cantonese became an important linguistic resource for expressing their emotions and highlighting their Hong Kong identity.

The studies outlined in this section show us that there are countless different factors that can affect language choice in online communication, including audience design, topic, social networks, gender, emotions and identity. Another factor that affects language use in general, including language choice, is a speaker's physical location and its spatial arrangement as well as the people and languages present in it. In the following section, I will discuss the relations between place and language practices.

2.4 Multilingualism, Online Language Use and Place

In this section, I take a brief look at some aspects of how physical space/location affects language use. In Section 2.4.1, I look at the role of place in online language use. Section 2.4.2 focuses on the connections between urban space and multilingualism and the concept of *metrolingualism*.

2.4.1 The Role of Place in Online Language Use

The interactions between physical space/place and language use in digital media is still a relatively under-researched area. However, there is an emerging stream of research in the field that has covered a diverse range of topics. According to Georgakopoulou (2015a), this research has focused, among other things, on how conceptions of place in online discussions shape and are shaped by identity, how the affordances and constraints of digital media affect discourse on place, how physical space affects online space and vice versa, and how localizing experiences is an important part of stance-taking (for examples of studies, see Cohen, 2015; Heyd and Honkanen, 2015; Georgakopoulou, 2015b).

An excellent example of this stream of research is a study by Heyd and Honkanen (2015). They examine linguistic representations of urban space on nairaland.com, an online discussion forum for members of the Nigerian digital diaspora. They point out that previous research on CMC (computer-mediated communication) has largely conceived of the Internet and cyberspace as purely abstract and virtual entities, with little or no connection to physical space. However, their study shows that explicit mentions and discussions of urban space are integral to the communications between members of the Nairaland forum. In particular, proximal deictic anchoring or "here in x" constructions emerge as ways of foregrounding a speaker's physical location. Members of the forum also express social meaning through the use of nonstandard toponyms such as *Chitown*, an affectionate name for Chicago, an important diasporic hub for the Nigerian diaspora. Discussion of different locations is common, and speakers take a wide variety of affective stances towards these places.

Heyd and Honkanen's findings indicate that place very much matters in online communication. It should be pointed out that these results are perhaps not as generalizable as the authors suggest. After all, nairaland.com is a forum that is directly related to physical locations (Nigeria as well as diasporic hubs) and discussions thereof, and other online spaces may well be more abstract and dislocated. However, Heyd and Honkanen also point out that on more recent social media platforms, place may play an even larger role than on traditional discussion forums. In my view, Instagram is a good example of this because of its geotagging

function as well as multimodal expressions of place through e.g. photographs, videos, captions and hashtags.

2.4.2 Urban Space and Multilingualism

Pennycook and Otsuji (2015) have developed the concept of *metrolingualism* to refer to "everyday multilingualism" (pp. 3) in urban spaces. Metrolingualism is very similar to *translanguaging*, a concept which criticizes the idea that languages are distinct, separate entities, and instead sees language and *languaging* as a process in which speakers draw from a repertoire of diverse semiotic resources, including different languages, registers and extralinguistic semiotic resources (Wei, 2017). According to Pennycook and Otsuji, the aim of metrolingualism is not to enumerate different codes or languages used in multilingual situations but rather to focus on the mobility of linguistic resources. However, whereas translanguaging focuses on individual speakers and their repertoires, metrolingualism instead foregrounds the interconnectedness of language practices and urban space.

In some ways, metrolingualism might seem incompatible with the present study. The idea of not enumerating languages is in contradiction with the study of language choice, where the languages being chosen between are indeed conceived of as at least somewhat separate entities. When using quantitative methods, the units being studied (in this case languages) must of course be defined. In addition, Pennycook and Otsuji (2015) stress that in multilingual situations, different languages often do not have distinct functions, whereas I, as well as many other researchers, would argue that often they do (including most of the studies cited in the previous sections, e.g. Leppänen and Nikula, 2007, and all of the studies in Section 2.3.1 that focus on language choice as a tool for designing one's audience). However, even though these ideas are controversial and contested, metrolingualism and language choice need not necessarily be irreconcilable approaches. We can acknowledge that, on the one hand, languages cannot be fully separated from each other and that the act of multilingual languaging is fluid, and on the other (as do in fact Pennycook and Otsuji as well as Wei) that speakers are nevertheless aware of different languages and their social and symbolic value, which affects their uses and functions.

In my view, the most salient aspect of metrolingualism is that (urban) space has a significant effect on language use (including language choice). Pennycook and Otsuji (2015) put forward the idea of *spatial repertoires*, meaning that people's language practices emerge locally through interaction with places and their spatial arrangements as well as the people and languages present in them. The concept of linguistic landscape is also relevant. As mentioned, Hiippala et al. (2018) characterize the Instagram posts geotagged to Helsinki Senate Square as part of the "virtual linguistic landscape" of the Senate Square. Pennycook and Otsuji also pay attention to the rhythms of the city, noting that linguistic landscapes change across different time frames. They also point out the layered and sedimented nature of texts and language in cities. In my view, the virtual linguistic landscape can be seen as another "layer" of language on top of the texts present in physical space as well as other current and historical linguistic practices.

3 Data

The data used in this study were collected by Hiippala et al. (2019). They were collected from Instagram, a highly popular social networking service whose main function is sharing photographs and videos along with captions and comments. The data consist of the captions of 117 418 Instagram posts by 74 051 users, uploaded over 1681 days (a little over four and a half years), between 4 July 2013 and 11 February 2018. Instagram allows posts to be geotagged to specific, pre-defined locations. The posts in this data set were collected from a 150-meter radius from the center of Helsinki Senate Square.

The likeliest country of origin for each Instagram user was determined based on where they have had the longest periods of activity. This allows me to study posts made by users who are very likely to be Finnish or at least live in Finland. The accuracy of this estimate was further reinforced by filtering out the lowest quarter of the users, i.e. those users whose country of origin was least certain. Users with eleven or fewer geotagged posts and whose longest period of activity was 44 days or less were excluded.

The language of each post was determined using automatic language identification (for a more detailed description of this process, see Hiippala et al., 2019). Automatic

language identification is of course not infallible, which is why posts shorter than 10 characters, as well as posts for which the likelihood of the language identification was fairly uncertain (less than 42.31%), were filtered out. This guarantees a very high accuracy rate of 91.9%. In this study, I am using the posts that were identified as being in either Finnish or English. After filtering for the country of origin of the users and the language of the posts in these ways, my final data set contains 16 987 posts with 19 304 orthographic sentences. Out of these sentences, 9 811 are in English and 9 493 in Finnish.

4 Methods and Qualitative Analysis

In this chapter, I discuss some ethical considerations as well as the methods used in this study. In Section 4.1, I outline some ethical issues related to protecting the Instagram users' anonymity and privacy and explain how I have attempted to solve these issues. In Section 4.2, I present the primary quantitative method used in this study, logistic regression. Section 4.3 explains how I determined the independent variables used in the logistic regression model: some were based on metadata provided by Instagram (time of posting), and some on analysis of the most frequent words in the data set as well as qualitative content analysis of random samples.

4.1 Ethical Considerations: Protecting Instagram Users' Anonymity

In order to illustrate my findings, I think it is important to include some examples from the data. However, this raises some important ethical issues concerning the anonymity of the Instagram users. While Instagram posts are technically considered public data and therefore no permission is required to use them for academic research, that does not mean it is always necessarily ethical to do so. Full posts or long enough quotations can often be easily traced back to individual users. With a platform like Instagram, this is especially problematic since most users regularly post pictures of themselves, and many even use their real names. While it could be argued that it is the Instagram users' own responsibility that they choose to make this information publicly available on the Internet, I would point out that they do not post on Instagram with the intention of having their language use singled out for academic analysis. People can have an expectation or perception of privacy even in what is

considered a public space, and we as researchers should take these expectations into consideration (Markham, 2012).

For this reason, I have taken some steps to anonymize the Instagram posts I am using as examples. Usernames have been replaced with pseudonyms (username1, username2 etc.). The posts used as examples have been slightly altered to prevent them from being easily traced back to users. For example, some lexical items have been replaced with synonyms. However, these alterations are small, and I have made an effort to keep the general meaning and tone of the posts as close as possible to the original. Some would of course argue that this is misrepresentation of users' language and therefore also unethical. However, in my view, protecting the users' anonymity and privacy takes priority. In addition, it is not my intention to analyze the language of these posts in any depth, but rather to provide them as examples that illustrate general tendencies in the data. Markham (2012) sees this kind of fabrication not as misconduct but as a useful and indeed ethical practice that is integral to protecting people's privacy when researching "public, archivable, searchable, and traceable" (pp. 336) online spaces.

4.2 Logistic Regression

The primary method of this MA thesis is logistic regression. Logistic regression is a quantitative statistical method used to model the probability of one of two binary, dichotomous outcomes with the help of independent variables or predictors. It is therefore well suited to modeling language choice between two languages (other regression models, such as linear regression, do not allow this). The binary dependent variable used in this study is the choice between Finnish (marked as 0) and English (marked as 1). The logistic regression model can include any number of independent variables, which allows me to study the effect of several different factors on language choice. The predictors used in this study are the time of posting, mention of the location and mention of popular events/holidays. The variables related to time of posting came out of metadata provided by Instagram, while the variables related to mention of the location and events were determined through qualitative content analysis (as explained in section 4.3). The probabilities of the language choice are expressed as odds ratios.

In this study, the logistic regression was performed using Python (version 3.7.0), a popular programming language for computational linguistics, data science, statistical analysis as well as many other purposes. Python has several libraries that are useful for this purpose. For this study, I used pandas (McKinney, 2010) to organize and handle the data, and statsmodels (Seabold and Perktold, 2010) to perform the logistic regression itself. I wrote and executed the Python script in Jupyter Notebook, a web application for coding. The results of the logistic regression are presented and analyzed in Chapter 5.

4.3 Determining the Independent Variables

In this section, I explain how I chose the variables included in the logistic regression model. This section is further divided into subsections. In Section 4.3.1, I discuss the variables related to time of posting, which is metadata provided by Instagram. In Section 4.3.2, I first outline the two ways in which I analyzed the data: analysis of the most frequent words in the data set and qualitative content analysis of random samples from the data. I then explain how I turned my observations from this analysis into variables and discuss some of the issues with the reliability of these variables.

4.3.1 Determining Variables Related to Time of Posting

Firstly, I wanted to look at the effect of external factors, in particular the time of posting, on language choice between Finnish and English. As noted by Pennycook and Otsuji (2015), linguistic landscapes tend to change along both cyclical and linear patterns, and I wanted to see if this was also the case in the virtual linguistic landscape of Helsinki Senate Square. The time of posting is metadata provided by Instagram.

I wanted to see whether the **hour**, **day**, and **month** of posting affect language choice, so these were turned into separate variables. In order to represent the cyclical nature of these time-related variables, they were each converted into a sine feature and a cosine feature. This conveys to the logistic regression model that when time is counted cyclically, points of time that seem to be numerically distant can actually be close to each other in time. For example, when using a 24 hour clock, 23:00 (11PM)

and 01:00 (1AM) numerically seem to always be 22 hours from each other, but because hours are counted cyclically, there is in fact only two hours between them.

I also thought it might be of interest to see whether different seasons have an effect, so I created variables for **summer** (June-August) and **winter** (November-December). These are binary variables that get a value of 0 or 1 based on whether a post was made during this time span.

As discussed in Section 2.2, Finnish people tend to use most English in their free time and for leisure activities (Leppänen et al., 2011). I wanted to see whether this is also the case in my data set, so created variables for **evening** (hours 18-23) and **weekend** (Saturday and Sunday), when most Finns are likely to have time off from work or school. These are also binary variables similar to the seasons.

Table 1 shows these variables and their explanations.

Variable	**Explanation**
hour_sin, hour_cos	hour of posting
day_sin, day_cos	weekday of posting
month_sin, month_cos	month of posting
summer	posted June to August
winter	posted November to January
evening	posted 6 PM to midnight
weekend	posted Saturday or Sunday

Table 1: Variables related to external factors and their explanations.

4.3.2 Determining Variables Related to Post Content

4.3.2.1 Analysis of Most Frequent Words

As well as Python, I also used AntConc, a freeware program for text analysis and corpus linguistics, to analyze my data set. AntConc's Word List function allowed me to produce a list of the most frequent words in the data. Studying this list gave me a good idea of some of the topics that occur most often in the data set.

Excluding stop words, by far the most common words in the data are related to the location in which the posts were made.

Token	Count
helsinki	13646
finland	5382
visithelsinki	2638
suomi	2335
senaatintori	2159

Table 2: The five most common words in the data excluding stop words. The total number of tokens in the data is 384238 and the number of types is 64981.

As we can see from Table 2, excluding stop words, the five most common words in the data set are all related to the city (‘helsinki’, ‘visithelsinki’), the country (‘finland’, ‘suomi’ [Finnish for Finland]) or the more specific location (‘senaatintori’ [Finnish for Senate Square]). These are all items which occur frequently in hashtags, which is why ‘visithelsinki’ is written without a space, and which in part explains their high frequency in the data set. Other words related to the location, such as ‘tuomiokirkko’ (‘Helsinki Cathedral’) and ‘cathedral’ are also common in the data. This is in line with findings by e.g. Heyd and Honkanen (2015) that online communication is by no means abstract and detached from physical space – on the contrary, in this data set, physical location is the most frequently discussed topic.

Another common theme emerging from studying the most frequent words is different events, festivals and holidays. These include holidays such as Christmas, New Year and Finnish Independence Day, as well as events such as Helsinki Christmas Market, Lux Helsinki (a light festival that takes places annually), Helsinki Pride (an LGBTIQ event culminating in a parade/demonstration) and Helsinki Midnight Run (an annual running event).

These two themes are clearly the most dominant, but other frequent topics also emerge. Many common words also relate to food and drink (e.g. ‘pizza’, ‘beer’), seasons and weather (e.g. ‘summer’, ‘cold’, ‘snow’) and travel (‘travelgram’, ‘trip’). It should be noted that many of the most commonly occurring words are very general in nature (e.g. ‘beautiful’, ‘time’, ‘love’, ‘like’, ‘new’, ‘people’) and it is therefore hard to say what topic they might relate to.

4.3.2.2 Qualitative Content Analysis of Random Samples

In addition to looking at the most frequent words, I also wanted to analyze the data in a more qualitative manner. With such a large data set (19 304 sentences), it is obviously not possible to read through all the captions. For this reason, I used the sample function in Pandas to take random samples from the data set. I took a sample of 300 sentences in English and 300 sentences in Finnish. I then used these samples for qualitative content analysis to identify common themes and other interesting elements that might be relevant to the study of language choice.

The themes I identified in these samples are very much in line with those emerging from the list of most frequent words. Many posts explicitly comment on the location, including the Senate Square and the Cathedral, the whole city, as well as the whole country. As illustrated by Example (1), often these elements can all occur within one post, and both within the body of the post and the hashtags:

(1) One of Helsinki's most well known buildings. What's your favorite? #Helsinki #finland #visitfinland #helsinkicathedral #building #famousbuildings #helsinkiphotos [username1, 2017]

As in the case of the most frequent words, events and holidays are also common topics in these samples – especially Christmas and the Helsinki Christmas Market, as illustrated by Example (2):

(2) Chatting with Santa at the Helsinki Christmas Market #Helsinki #Christmas #travel #Santa #ig_helsinki [username2, 2015]

Food, drink, weather and travel are also among the most common talking points:

(3) Maybe the best pizza in town [username3, 2017]

Some themes in these samples that are not so apparent in the list of most frequent words include work/job related posts and commercial posts by companies. These can be found especially in the Finnish sample, but also to some extent in the English one.

4.3.2.3 Turning Qualitative Observations into Variables

My observations of the most frequent words and the random samples support each other. The two biggest and most interesting themes emerging from both are 1) discussion of the location and 2) different events and holidays. As well as being

frequent in the data, I would argue that these topics are also relevant for the study of language choice. As a popular tourist attraction, it is possible that the Senate Square is viewed as a highly "international" place, which might encourage the use of English over Finnish. In general, English is more frequently used in urban areas of Finland than in rural ones (Leppänen et al. 2011), so discussion of an urban place might also encourage the use of English. As noted by e.g. Pennycook and Otsuji (2015) as well as Heyd and Honkanen (2015), physical location in general and urban space in particular has a significant effect on people's language practices, and I would argue also their language choices. As previously mentioned, English is also used most frequently during Finnish people's free time and for leisure activities (Leppänen et al. 2011), so I thought it would be interesting to see if discussion of holidays and free-time events influences language choice.

I used the same method for all the variables related to the content of the Instagram posts. For each variable, I looked for instances of various keywords/phrases in each Instagram caption in the data set. If one or more of these key items was present in the post, the variable got a value of 1, and if not, a value of 0. These items can occur either as separate words or embedded e.g. in hashtags, which are written without spaces.

In the case of the location, I thought it would be of interest to study the effect of the specific location (the Senate Square and Helsinki Cathedral), the whole city (Helsinki), as well as the whole country (Finland). **location_mentioned** checks for mentions of the following items: 'senate square', 'cathedral', 'senaatintori' (Finnish for Senate Square) and 'tuomiokirkko' (Finnish for Helsinki Cathedral). **city_mentioned** checks for 'helsinki' and 'stadi' (a common Finnish slang term for Helsinki). **country_mentioned** checks for 'finland' and 'suomi' (the Finnish name for Finland).

In the case of the holidays and events, I picked some of the ones that occur most frequently in the data and that could be turned into keywords/phrases relatively straightforwardly. I combined all the events into a single variable, **event_mentioned**. The holidays and events contained in this variable, as well as the key items for each of them, are as follow:

- Christmas: ‘christmas’, ‘xmas’, ‘joulu’ (Finnish for Christmas)
- Helsinki Christmas Market (and possibly other similar events): ‘market’, ‘markkinat’ (Finnish for market/fair)
- Lux Helsinki, an annual light festival: ‘lux’
- Helsinki Pride, an event for sexual and gender minority rights: ‘pride’
- Finnish Independence Day: ‘independence’, ‘itsenäisyys’ (Finnish for independence)
- New Year: ‘new year’, ‘uusi vuosi’ (Finnish for new year), ‘uutta vuotta’ (Finnish for new year in an inflected form that often occurs in greetings)
- Helsinki Midnight Run, an annual sports event: ‘midnight run’, ‘midnightrun’

Table 3 shows all the variables related to the content of the posts as well as the keywords/phrases used.

Variable	**Keywords/phrases**
location_mentioned	‘senaatintori’, ‘tuomiokirkko’, ‘senate square’, ‘cathedral’
city_mentioned	‘helsinki’, ‘stadi’
country_mentioned	‘finland’, ‘suomi’
event_mentioned	‘christmas’, ‘xmas’, ‘joulu’, ‘market’, ‘markkinat’, ‘lux’, ‘pride’, ‘independence’, ‘itsenäisyys’, ‘new year’, ‘uusi vuosi’, ‘uutta vuotta’, ‘midnight run’, ‘midnightrun’

Table 3: Variables related to post content with keywords/phrases.

It should be noted that these variables are by no means perfect. Obviously, the keywords/phrases do not capture every instance of reference to these places and events. For example, many other formulations can be used to refer to Helsinki, such as “my hometown” or “the capital city”, and it is impossible to look for all of these. The same goes for all these variables, but I think I have included the most common and important words used to refer to the places and events. Because the geotagging works on a radius of 150 meters, some posts refer to places around the Senate Square, such as the University of Helsinki and different restaurants, and these obviously fall outside my analysis unless they also mention the Square or the Cathedral. However, this is not really an issue for me since the Square and the Cathedral are the locations I wish to focus on.

Another problem is the highly inflectional nature of the Finnish language. Searching for the nominative form of a word does not capture all of its uses; for example, the

word ‘Helsinki’ sometimes appears as ‘Helsingin’ or ‘Helsingissä’. Fortunately, instances in which the stem of a word does not change when the word is inflected are captured, since the keywords/phrases can be embedded as well as separate. Getting more reliable variables would require additional processing using language technology that captures all the possible grammatical forms of each keyword/phrase, but this is unfortunately beyond the scope of this study. Despite these issues, I think my results will nevertheless give a fairly reliable picture of how discussion of certain topics (the physical location and events/holidays) affects language choice for Finnish Instagram users. However, it is important to keep in mind that these issues exist, and that further research is needed to confirm my results.

5 Results for Logistic Regression and Analysis

In this chapter, I present and discuss the results of the logistic regression. In section 5.1, I analyze the results related to the time of posting, and in section 5.2, the results related to the content of the Instagram posts, namely how the mention of the physical location of posting and the mention of different events/holidays affect language choice for Finnish Instagram users. Table 4 shows a summary of all the results (but the relevant results are presented again in each subsection).

Independent variable	**Odds ratio**
hour_sin	0.97
day_sin	0.87
month_sin	1.03
hour_cos	1.31
day_cos	1.07
month_cos	1.35
weekend	0.92
evening	0.81
summer	0.75
winter	0.77
location_mentioned	1.02
city_mentioned	1.61
country_mentioned	2.43
event_mentioned	0.66

Table 4: Summary of all the results of the logistic regression. The table shows the odds ratio for each independent variable.

As mentioned in Section 4.2, the results of logistic regression are expressed as odds ratios. The odds ratio of an independent variable shows how a one unit increase in that variable, holding all other variables constant, affects the odds of the dependent variable (in this case the choice of language). An odds ratio of one indicates that there is no change in the odds, i.e. the variable has no effect on language choice. An odds ratio above one indicates that the odds of choosing English increase (and therefore the odds of choosing Finnish decrease) when the value of the variable in question increases. An odds ratio below one indicates that the odds of choosing English decrease (and therefore the odds of choosing Finnish increase) when the value of the variable increases. For example, if we have a binary independent variable that gets a value of either 0 or 1 based on whether a post contains a specific word, and the odds ratio for this variable is 1.5, this means that the inclusion of this word in a post (an increase in value from 0 to 1) increases the odds of the post being in English by a factor of 1.5.

5.1 Time of Posting

Independent variable	Odds ratio
weekend	0.92
evening	0.81
summer	0.75
winter	0.77

Table 5: Results for variables related to time of posting.

The results of the variables which model time in a cyclical manner are hard to interpret. However, the important thing to note is that the results for logistic regression strongly indicate that the time of posting has little to no effect on language choice. An alternative version of the logistic regression model was also performed with all of the time-related variables removed. The results of this model were similar to the one with time of posting taken into account; this shows that including time-related information in the model does not change the results and therefore that time of posting is unlikely to affect the choice of language.

As Table 5 shows, the non-cyclical time-related variables also seem to have little effect on language choice. As mentioned in Section 4.3.1, my hypothesis was that using English would be more likely during evenings and weekends as well as in the summer, when people are likely to have time off from work or school. However, the

logistic regression model does not support this. In fact, it seems that odds of choosing English decrease slightly during these times. This is somewhat surprising given the earlier findings by e.g. Leppänen et al. (2011) that English in Finland is largely associated with free time and fun. I also would have expected Finnish people in Helsinki to use more English in the summer, because there is an increased presence of tourists in the summer, especially in places like the Senate Square. This is also something that is often commented on in Instagram posts:

(4) Spotting my tourist hat is a certain sign of summer #summer #helsinki #tourist #ootd [username4, 2017]

In example (4), the connection between summer and tourism is made both in the body of the caption as well as the hashtags. In general, tourism and "playing tourists" are quite common themes in Finnish Instagram users' posts, which I would have expected to also affect language choice. However, my results indicate that there is much less correlation between the time or season of posting and language choice than hypothesized. Perhaps the uses and functions of English have already changed rather drastically from when Leppänen et al. carried out their study (2011, survey conducted in 2007) so that the use of English is no longer so clearly concentrated in any specific time.

5.2 Post Content

	coef	**std err**	**z**	**P > | z |**	**[0.025**	**0.975]**
location_mentioned	0.0299	0.040	0.744	0.457	-0.049	0.109
city_mentioned	0.4955	0.032	15.589	0.000	0.433	0.558
country_mentioned	0.8615	0.039	22.163	0.000	0.785	0.938
event_mentioned	-0.3280	0.034	-9.568	0.000	-0.395	-0.261
intercept	-0.3458	0.023	-14.970	0.000	-0.391	-0.301

Table 7: Detailed summary of results for variables related to post content.

In this section, I present the results for the variables related to the content of the Instagram posts. Section 5.2.1 concerns the mention of the location of posting, and Section 5.2.2 the mention of events or holidays. Table 7 shows a more detailed summary of the logistic regression results related to post content. The column labeled '**coef**' contains the coefficients for the variables; these have been converted into odds ratios in the other tables. '**std err**' stands for standard error. This is relatively low for all the variables. The **z**-value is the coefficient divided by its standard error. The

further away a z-value is from zero, the more the variable matters. The next column presents the *p*-value for each variable, which indicates whether the value for the coefficient is statistically significant. Most notably, the *p*-value for the variable **location_mentioned** is 0.457, which means that it is not statistically significant ($p < 0.05$ is usually taken as a cut-off value for statistical significance). The *p*-value for the other variables, on the other hand, is <0.000. In other words, it is highly unlikely that these results are coincidental. The last two columns show the variable's 95% confidence interval, meaning that there is a 95% chance that the coefficient value is between these two values.

5.2.1 Mention of Location

Independent variable	Odds ratio
location_mentioned	1.02 (not statistically significant)
city_mentioned	1.61
country_mentioned	2.43

Table 8: Results for variables related to the physical location.

As discussed in Section 4.3.2, the location of posting is often mentioned in Instagram posts, and it is the most frequent talking point/topic in the data. Example 5 is a very typical caption for this data set:

(5) Helsinki Cathedral #cathedral #helsinki #finland #snow [username5, 2017]

Here, the location is mentioned both in the tags and the body of the caption. This example includes mention of the specific location, the city as well as the whole country, but many posts only contain one or two of these.

Table 8 shows the results related to the mention of location. My hypothesis was that mentioning the specific location, in other words the Senate Square or the Helsinki Cathedral, would make the use of English more likely since the Square and Cathedral are popular tourist attractions and might therefore be seen as highly "international" places. However, as previously discussed, the results for the variable **location_mentioned** are not statistically significant, so this hypothesis could not be confirmed.

Mention of the city has a significant effect, increasing the odds of choosing English by a factor of approximately 1.6. A possible explanation for this could be that use of

English in Finland is associated with urban areas (Leppänen et al. 2011), and explicit mention of a city and therefore awareness of being in an urban area might affect language choice. As pointed out by Pennycook and Otsuji (2015) as well as Heyd and Honkanen (2015), (urban) spaces have an effect on people's language practices. It seems that English plays an important role in the *spatial repertoire* (Pennycook and Otsuji, 2015) of the Senate Square.

Mention of the country is by far the most significant factor affecting language choice in this study. It increases the odds of choosing English by a factor of approximately 2.4. This is a rather surprising result which might be explained by issues of audience design (Bell, 1984; see Section 2.3). Posts in which the country is explicitly mentioned might be directed at a more international audience and therefore written in English. When directing a post to a Finnish audience, Instagram users might not feel the need to specify that they are in Finland because this is the default expectation. Most Finnish people are also likely to know that the Senate Square and the Cathedral are located in Finland and therefore do not need to be given this information, but this might not be the case for an international audience. On the other hand, I would have expected a lot of Finland-specific posts as well as patriotic posts discussing Finland in Finnish, but perhaps these only make up a small minority of the posts in which Finland is mentioned.

Another possible reason why Helsinki and Finland are so often mentioned in posts directed at an international audience is that the Senate Square and especially the Cathedral may function as symbols of Finland and Finnishness. As mentioned, the Square and Cathedral are some of the most popular tourist attractions in Finland. In a study on churches in the tourism images of Helsinki, Jokela (2014) states that the Cathedral is depicted in two thirds of all the images and characterizes the Cathedral as "Helsinki's modest equivalent of the Eiffel Tower" (pp. 624). According to Jokela, the Cathedral has been a "pervasive symbol in tourism images and in the underlying identity-political discourses through which hegemonic conceptions of Finland and Helsinki have been constructed" (2014, pp. 253). The Cathedral has taken on various different meanings throughout its history. Being an important Lutheran church, it has been a symbol of the West (as opposed to Orthodox Christianity's association with the East and especially Russia). During the interwar

period, it became associated with hegemonic conceptions of Finnishness, such as the values of "home, patria and religion" (Jokela, 2014, pp. 626). Its white façade has been tied to Finland's image as a "pure and white" nation as well snow and the exotic North (Jokela, 2014, pp. 626). According to Jokela, the Cathedral has also symbolized the connection between Lutheranism and the Finnish state, and many state-related events have been held there.

Thus, by associating the Senate Square and Helsinki Cathedral with Finland and Finnishness (through tags or in the body of the caption itself) and posting about it in English, Finnish Instagram users perhaps wish to project a certain (hegemonic) image of Finland to the rest of the world. Some of the associations observed by Jokela (2014) certainly come up in the present data. For example, the whiteness of the Cathedral is underlined by many users:

(6) The White Cathedral #white #cathedral #helsinki #finland #snow [username6, 2017]

In example (6), username6 has chosen to use capital letters in the phrase 'The White Cathedral', suggesting that it is used as an established description/nickname that carries specific meanings and not just an ad hoc comment on the color.

5.2.2 Mention of Events/Holidays

Independent variable	Odds ratio
event_mentioned	0.66

Table 9: Odds ratio for mention of holidays/events.

As discussed in Section 4.3.2, mentions of popular holidays and events that take place in the Senate Square occur often in the data. Examples (7) and (8), in which a simple comment is made about the event in question, are very typical in my data set.

(7) Experiencing Helsinki Christmas market #finland [username7, 2017]

(8) Happy Pride to everyone in Helsinki! #pride #pridehelsinki2016 #helsinki [username8, 2016]

I chose some of the most frequently discussed events and holidays (Christmas, New Year, Finnish Independence Day, Helsinki Christmas Market, Lux Helsinki, Helsinki Pride, Helsinki Midnight Run) and combined them into a single variable.

The mention of these events has a moderate effect on language choice. Surprisingly, Finnish Instagram users are less likely to use English when discussing events. This, too, goes somewhat against my hypothesis, since it could be argued that these events all relate to free time and fun in some way, yet they are more likely to be talked about in Finnish than in English. One explaining factor could be that some of these events may be seen as "local" or "Finnish" and therefore as not having much international relevance. It is possible that people tend to use the local language for local topics. Helsinki Christmas Market, Lux Helsinki and Helsinki Midnight Run especially might be events that are seen as "local" since they are specific to Helsinki and perhaps not so relevant for people who cannot attend them.

The Finnish Independence Day, in turn, might be seen as "inherently Finnish" and therefore posting about it in Finnish might feel more natural. Christmas is also largely viewed as a traditional Finnish holiday. On the other hand, both Christmas and New Year are celebrated in many different countries all around the world, so I would have expected Christmas and New Year's greetings to be mainly in English. Pride is also an event that takes place in many different countries and has its origins in the US, and the struggles of LGBTIQ people are definitely not unique to Finland. I would therefore argue that these events are more "global" or "international" in nature and would have expected more English. It would be very interesting to study the effect of these events individually and see if there are differences between them, but that is unfortunately beyond the scope of this MA thesis.

6 Discussion

6.1 Summary of Main Results

The results of logistic regression indicate that the time of posting has no effect on language choice. Performing the logistic regression with or without the time-related variables does not change the rest of the results. Contrary to my hypothesis, the odds of choosing English do not increase during weekends or evenings or in the summer; in fact, the odds of choosing Finnish slightly increase. Mention of the physical location, on the other hand, has a very significant effect – especially the mention Finland, which surprisingly increases the odds of choosing English by a factor of 2.4. The mention of Helsinki also has a major effect, increasing the odds of English by a

factor of 1.6. The mention of different holidays or events (Christmas, New Year, Finnish Independence Day, Helsinki Christmas Market, Lux Helsinki, Helsinki Pride, Helsinki Midnight Run) in a post also has an effect on language choice, but unlike the mention of the location, mention of these events decreases the odds of choosing English. In other words, Instagram users are much more likely to make posts about these events in Finnish.

6.2 Comparisons with Earlier Research

As discussed by Hiippala et al. (2019), it is notable that around half of the posts made by Finnish people in this data set are in English. This finding was one of the motivating factors for the present study. Finnish people's increasing use of English in general has been noted by several researchers. Leppänen and Nikula (2007) as well as Taavitsainen and Pahta (2008) argue that English is becoming a more domesticated language in Finland, used in intranational contexts as well as international ones, and an increasingly natural part of Finns' linguistic repertoires. A large national survey (Leppänen et al., 2011) found that more than a quarter of Finns consider using English as natural as using their mother tongue. Almost half of the respondents also indicated that they do not even notice that they are using English when writing (for spoken interaction, the figure was even higher). Therefore, the high degree of English use in my Instagram data could simply be explained by the fact that Finnish people are using more and more English in general and making the language their own. On the other hand, Leppänen et al. (2011) also found that Finns use more English on the Internet than in offline interactions. In addition, English in Finland is associated with urban areas (Leppänen et al., 2011), so the fact that the data was gathered in the center of Helsinki might also be significant. Physical space influences people's language practices (e.g. Heyd and Honkanen, 2015), and since the Senate Square is an "international" place frequented by tourists, it could be that English plays an important role in the *spatial repertoire* (Pennycook and Otsuji, 2015) of the Square. In the light of these observations, perhaps the prevalence of English in the data set should not be all that surprising.

Another explaining factor for the frequent use of English could be audience design (Bell, 1984; see Section 2.3). Instagram users may wish to maximize their audience (Androutsopoulos, 2014) by using English in order to gain more followers or to

simply appeal to a wider, more international audience. This audience design strategy might also be due to more practical reasons, e.g. out of consideration for friends who do not speak Finnish.

The results for logistic regression can also to some extent be explained by earlier research. As mentioned, English is most frequently used in urban areas of Finland (Leppänen et al., 2011), which might in part explain why mention of Helsinki increases the odds of choosing English. However, the finding that mentioning Finland in a post dramatically increases the odds of choosing English is rather surprising in the light of previous research. In fact, it goes against the observation made by some researchers (e.g. Hinrichs, 2016) that local topics tend to be discussed in local languages. Instead, it might be explained in part by issues of audience design and in part by the Cathedral's connection with certain conceptions of Finland and Finnishness (see Jokela, 2014) that Finns wish to project to the rest of the world.

Unlike the mention of Finland, the mention of holidays and events increases the odds of choosing Finnish, which might be explained by the connection between local topic and local language. At least some of these events (especially those that are specific to Finland) may be viewed as local or inherently Finnish in nature, and therefore not of relevance for an international audience. However, some of the holidays and events are arguably more internationally relevant, which complicates this interpretation.

Another very surprising finding in the light of previous research is that the time of posting does not seem to have any effect on language choice. According to Pennycook and Otsuji (2015), the rhythms and cycles of a city have a significant effect on the spatial repertoires and linguistic landscapes of a city, but this is not reflected in my results. As mentioned, English has also been associated with free time and fun in Finland (Leppänen et al., 2011), but my results do not show any increase in the use of English during most people's free time (e.g. weekends, evenings and summers) relative to Finnish.

I would argue that the present study adds to previous research in several important ways. Firstly, at the time of writing, research on English in Finland has so far not addressed language use on SNSs, even though Facebook has now been around for as long as 16 years, and Twitter for 14. Even Instagram, which is a relative newcomer,

has existed for 9 years. SNSs have become an important part of many Finns' daily lives and language practices. I therefore think that it is crucial for linguists not to leave out these platforms if we wish to have a complete picture of Finnish people's language use. Secondly, unlike Facebook, which has been extensively studied, Instagram is in general an under-researched platform, especially when it comes to language choice. As mentioned in Section 2.3, to my knowledge, there is only one previous study on language choice on Instagram. Thirdly, the effects of the time of posting and the topic of a post on language choice in SNSs have not been previously studied. Previous research focuses heavily on questions of audience and identity, and although many studies mention the effect of topic on language choice, none of them focus on it in any detail.

The present study is also a rare example of a quantitative approach to language choice on SNSs. Most of the previous studies are qualitative in nature, and although qualitative research is of course just as valuable as quantitative research, I would argue that both are needed in order to get a complete picture of the language situation. Finally, this study also adds to the slowly emerging stream of research on how place, space and location affect language practices online.

6.3 Limitations

Some of the limitations of this study have already been discussed in previous chapters. As discussed in section 4.3, the variables used in the logistic regression are not perfect. The variables related to the content of the posts do not capture every instance of reference to the locations and the events studied. The inflectional nature of the Finnish language means that some of the grammatical forms of certain words are not captured. Getting more reliable variables would therefore require language technological analysis that captures every form of a Finnish word.

An important limitation is that my results cannot be generalized to all Finnish Instagram users and certainly not all Finnish people. The results are specific to the location in which the data were gathered as well as the particular SNS. It would be very interesting to conduct a similar study on Instagram data gathered from a different geotagged location and compare the results.

This study only considers a few of the many possible factors affecting language choice: time of posting and mentions of locations and events. Of course, this is inevitable because there is only so much that can be studied in one MA thesis. There are many other topics that I could have chosen which no doubt affect language choice. Other factors not considered in this study or considered only briefly include audience design (Bell, 1984; see Section 2.3), identity, user biographies, technological factors, the characteristics of individual social networks and so on.

Unfortunately, this study also does not consider code-switching or related phenomena. Because I wished to study the factors affecting choice between Finnish and English, the data set only includes posts which are *either* in Finnish *or* in English. Of course, this leaves out the cases in which a user chose to use both or include other languages. In a way, the study of language choice also ends up supporting a somewhat static view of language, in which different languages can be neatly separated from each other. This view has been criticized by many (e.g. Wei, 2017 and Pennycook and Otsuji, 2015). However, I contend that perceived differences between languages, however socially constructed they may be, remain important for speakers, which affects how they use these languages.

Since Instagram is a platform designed primarily for sharing photographs and short videos, it could be argued that studying only the captions leaves out an important part of the visual context in which they appear. In a study using quantitative methods and a large data set, this is somewhat inevitable, but a qualitative study using multimodal analysis would also be interesting.

6.4 Suggestions for Further Research

A single MA thesis can only scratch the surface of such a broad and under-researched topic. There is ample opportunity for further research in the area of language choice in SNSs. As already mentioned, it would be very interesting to study the effect of the physical location by gathering Instagram data from another geotagged location, perhaps in a suburban or rural area or another, smaller city in Finland, and comparing the results to the present study. Another option would be to conduct a more comprehensive version of the present study using deeper language technological analysis and separate variables for all the events and holidays studied.

The present data set could also be used to study how the mention of e.g. food or the weather affects language choice, although these topics are much harder to turn into variables.

As mentioned previously, in the field of linguistics, Instagram is a very under-researched platform, which offers a lot of scope for research. The same frameworks that have previously been applied to the study of language choice on Facebook and Twitter could also be applied to Instagram. These include for example audience design (Bell, 1984; see Section 2.3) and network factors (Eleta and Golbeck, 2014). Although the present study uses mixed methods, the primary method is quantitative, so there is a lot of space for closer, qualitative analysis of Instagram posts. Ideally, these studies could use multimodal analysis on photographs, captions, emojis, the layout of the Instagram app etc., since these are all important components of the platform.

A qualitative study could also focus on code-switching/code-mixing and related phenomena. More studies could also be done on language choice in hashtags (see Lee and Chau, 2018), as these are also an important element of Instagram. Another interesting research opportunity would be to compare language choice on Instagram to language choice on another SNS such as Facebook or Twitter, perhaps using SIDE theory (see Birnie-Smith, 2015). The present study and other similar studies could also be complemented by surveys or even interviews asking Instagram users about their language choices on Instagram and what they consider to be some of the reasons for these choices, although the questionable reliability and generalizability of self-reporting must be kept in mind.

7 Conclusion

In this MA thesis, I have studied the effect of several different factors on the language choices of Finnish Instagram users in Helsinki Senate Square. This was done using logistic regression, a quantitative method for measuring the effect of multiple variables on a binary outcome (in this case, the choice of either Finnish or English). These variables were chosen by studying the list of most frequent words in the data set as well as using qualitative content analysis on random samples of the data. The data set used was a large collection of Instagram posts gathered by

Hiippala et al. (2019). The results of logistic regression indicate that in this data set, the time of posting has little to no effect on language choice. Mentioning the location of posting in a post, on the other hand, has a much more significant effect. Somewhat surprisingly, mentioning Finland significantly increases the odds of choosing English. Also surprising is that mentioning certain popular events or holidays decreases the odds of choosing English.

Of course, these results represent only a few of all the possible factors affecting language choice for Finnish Instagram users. However, they are a valuable starting point in filling a rather large gap in the body of research on English in Finland, which has previously not addressed any SNSs. In addition, Instagram is an under-researched platform, especially when it comes to the study of language choice. Previous studies on language choice in SNSs have also tended to be qualitative and use relatively small data sets. This study has demonstrated the usefulness of logistic regression as a method for studying language choice in larger sets of SNS data. Furthermore, the combination of logistic regression with qualitative analysis gives us new insights into language choice. Online language choice in general and especially the role of English in multilingual contexts are important topics which deserve to be studied. I therefore hope this MA thesis can serve as a stepping stone for both broader and deeper studies into these and related issues.

References

Androutsopoulos, J., 2007. Language Choice and Code Switching in German-Based Diasporic Web Forums. In: B. Danet and S. Herring, eds., *The multilingual Internet*, 1st ed. Oxford: Oxford University Press, pp. 341-362.

Androutsopoulos, J., 2014. Languaging when contexts collapse: Audience design in social networking. *Discourse, Context & Media*, *4-5*, pp. 62-73. doi: 10.1016/j.dcm.2014.08.006

Bell, A., 1984. Language style as audience design. *Language In Society*, 13(02), 145-204. doi: 10.1017/s004740450001037x

Birnie-Smith, J., 2015. Ethnic identity and language choice across online forums. *International Journal Of Multilingualism*, 13(2), pp. 165-183. doi: 10.1080/14790718.2015.1078806

Coats, S., 2019. Language choice and gender in a Nordic social media corpus. *Nordic Journal of Linguistics*, 42, pp. 31-55. doi: 10.1017/S0332586519000039

Cohen, L., 2015. World attending in interaction: Multitasking, spatializing, narrativizing with mobile devices and Tinder. *Discourse, Context & Media*, 9, pp. 46-54.

Cunliffe, D., Morris, D., & Prys, C., 2013. Young Bilinguals' Language Behaviour in Social Networking Sites: The Use of Welsh on Facebook. *Journal Of Computer-Mediated Communication*, 18(3), pp. 339-361. doi: 10.1111/jcc4.12010

Eleta, I. and Golbeck, J., 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers In Human Behavior*, 41, pp. 424-432. doi: 10.1016/j.chb.2014.05.005

Georgakopoulou, A., 2015a. Introduction: Communicating time and place on digital media—Multi-layered temporalities & (Re)localizations. *Discourse, Context & Media*, 9, pp. 1-4.

Georgakopoulou, A., 2015b. Sharing as rescripting: Place manipulations on YouTube between narrative and social media affordances. *Discourse, Context & Media*, 9, pp. 64-72.

Heikkinen, M., 2018. Suomen kieli on uhattuna, tällaisia keinoja poliitikot tarjoavat – Testaa, ymmärrätkö uussuomen lainasanahirviöitä. *Helsingin Sanomat*, [online]. Available at: https://www.hs.fi/kulttuuri/art-2000005931218.html?fbclid=IwAR0ICagf9KrD6lN_CfuW2qX4Pkqn36kYoFnToYrevRQi34b53x6uGRmgMkw [Accessed 9 Jan 2019].

Heyd, T. and Honkanen, M., 2015. From Naija to Chitown: The New African Diaspora and digital representations of place. *Discourse, Context & Media*, 9, pp. 14-23.

Hiippala, T., Hausmann, A., Tenkanen, H. and Toivonen, T., 2019. Exploring the linguistic landscape of geotagged social media content in urban environments. *Digital Scholarship in the Humanities*, 34(2), pp. 290-309.

Hinrichs, L., 2016. Modular Repertoires in English-Using Social Networks: A Study of Language Choice in the Networks of Adult Facebook Users. In: L. Squires, ed., *English in Computer-Mediated Communication: Variation, Representation, and Change*, 1st ed. Berlin: De Gruyter, pp. 17-42.

Jokela, S., 2014. Tourism and identity politics in the Helsinki churchscape. *Tourism Geographies*, 16(2), pp. 252-269.

Kääntä, L., Jauni, H., Leppänen, S., Peuronen, S. and Paakkinen, T. (2013). Learning English Through Social Interaction: The Case of Big Brother 2006, Finland. *The Modern Language Journal*, 97(2), pp. 340-359.

Laitinen, M., 2014. 630 kilometres by bicycle: observations of English in urban and rural Finland. *International Journal of the Sociology of Language*, 2014(228), pp. 55-77.

Laitinen, M., Lundberg, J., Levin, M. and Martins, R., 2018. The Nordic Tweet Stream: A Dynamic Real-Time Monitor Corpus of Big and Rich Language Data. In: *DHN 2018 Digital Humanities in the Nordic Countries 3rd Conference*. Helsinki, Finland, March 7-9, 2018, pp. 349-362.

Lee, C. and Chau, D., 2018. Language as pride, love, and hate: Archiving emotions through multilingual Instagram hashtags. *Discourse, Context & Media*, 22, pp. 21-29. doi: 10.1016/j.dcm.2017.06.002

Lee, C., 2014. Language choice and self-presentation in social media: the case of university students in Hong Kong. In: P. Seargeant and C. Tagg, eds., *The Language of Social Media: Identity and Community on the Internet*, 1st ed. London: Palgrave Macmillan, pp. 91-111.

Leppänen, S. and Nikula, T., 2007. Diverse uses of English in Finnish society: Discourse-pragmatic insights into media, educational and business contexts. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, 26(4), pp. 333-380.

Leppänen, S., 2007. Youth language in media contexts: insights into the functions of English in Finland. *World Englishes*, 26(2), pp. 149-169.

Leppänen, S., Pitkänen-Huhta, A., Nikula, T., Kytölä, S., Törmäkangas, T., Nissinen, K., Kääntä, L., Räisänen, T., Laitinen, M., Pahta, P., Koskela, H., Lähdesmäki, S. and Jousmäki, H., 2011. National survey on the English language in Finland: Uses, meanings and attitudes. *Studies in Variation, Contacts and Change in English* Vol. 5, [online] Available at: <http://www.helsinki.fi/varieng/series/volumes/05/> [Accessed 3 October 2018].

Leppänen, S., Pitkänen-Huhta, A., Piirainen-Marsh, A., Nikula, T. and Peuronen, S., 2009. Young People's Translocal New Media Uses: A Multiperspective Analysis Of Language Choice And Heteroglossia. *Journal Of Computer-Mediated Communication*, 14(4), pp. 1080-1107. doi: 10.1111/j.1083-6101.2009.01482.x

Markham, A., 2012. Fabrication as Ethical Practice. *Information, Communication & Society*, 15(3), pp. 334-353. doi: 10.1080/1369118x.2011.641993

McCay-Peet, L. and Quan-Haase, A., 2017. What is Social Media and What Questions Can Social Media Research Help Us Answer? In: L. Sloan and A. Quan-Haase, eds., *The SAGE Handbook of Social Media Research Methods*, 1st ed. London: SAGE Publications Ltd, pp. 13-26.

McKinney, W., 2010. Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference (SciPy 2010)*. [online] Austin, Texas, pp. 51-56. Available at: https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf [Accessed 14 Jun 2019]

Page, R., Barton, D., Unger, J. and Zappavigna, M., 2014. *Researching Language and Social Media*. New York: Routledge.

Pennycook, A. and Otsuji, E., 2015. *Metrolingualism – Language in the city*. London: Routledge.

Piirainen-Marsh, A., 2010. Bilingual practices and the social organisation of video gaming activities. *Journal of Pragmatics*, 42(11), pp. 3012-3030.

Seabold, S. and Perktold, J., 2010. Statsmodels: Econometric and Statistical Modeling with Python. In: *Proceedings of the 9th Python in Science Conference (SciPy 2010)*. [online] Austin, Texas, pp. 57-61. Available at: http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf [Accessed 14 Jun 2019]

Seargeant, P., Tagg, C. and Ngampramuan, W., 2012. Language choice and addressivity strategies in Thai-English social network interactions. *Journal Of Sociolinguistics*, 16(4), pp. 510-531. doi: 10.1111/j.1467-9841.2012.00540.x

Soler-Adillon, J. and Freixa, P., 2017. Wikipedia access and contribution: Language choice in multilingual communities. A case study. *Anàlisi*, (57), pp. 63-80. doi: 10.5565/rev/analisi.3109

Taavitsainen, I. and Pahta, P., 2003. English in Finland: globalisation, language awareness and questions of identity. *English Today*, 19(4), pp. 3-15.

Taavitsainen, I. and Pahta, P., 2008. From global language use to local meanings: English in Finnish public discourse. *English Today*, 24(03), pp. 25-38.

Tagg, C. and Seargeant, P., 2014. Audience design and language choice in the construction and maintenance of translocal communities on social network sites. In: P. Seargeant and C. Tagg, eds., *The Language of Social Media: Identity and Community on the Internet*, 1st ed. London: Palgrave Macmillan, pp. 161-185.

Wei, L., 2017. Translanguaging as a Practical Theory of Language. *Applied Linguistics*, 39(2), pp. 261-261.

Zappavigna, M., 2012. *Discourse of Twitter and social media*. London: Continuum International Pub. Group.